Muddled thinking in UK policy evaluation (and why we should try to change it)

Dr Andi Fugard, IFF Research

University of Cambridge MPP Forum
26 Feb 2026

The plan

IFF Research

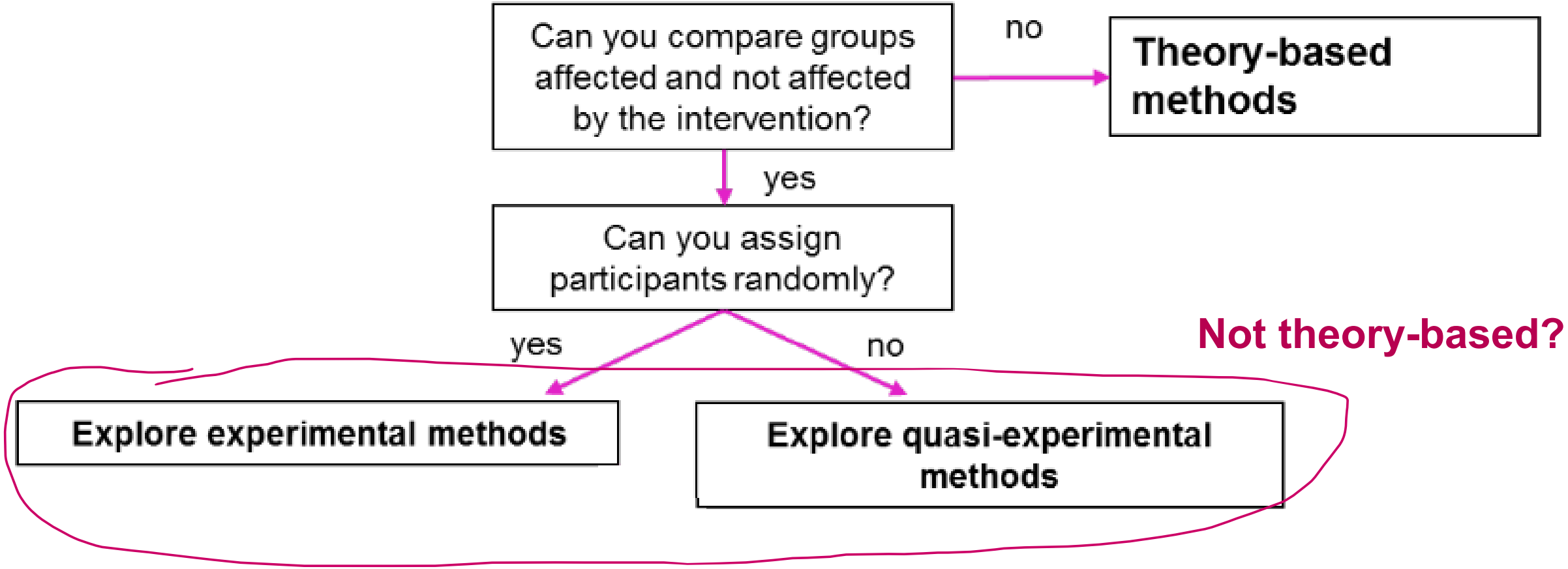
Examples of how UK policy evaluation gets muddled

How you can lend a hand to improve things

Why it matters

Muddle 1:
“theory-based” evaluation

“Theory-based methods” (Magenta Book)



The muddle (Magenta, p. 43)

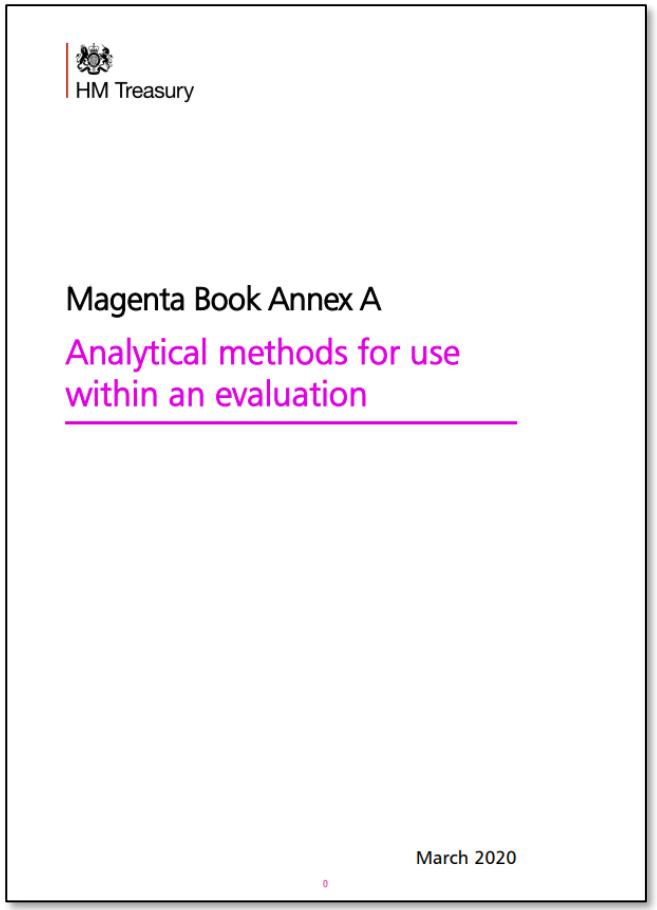
IFF Research

"Theory-based methods can be used to investigate net impacts by exploring the causal chains thought to bring about change by an intervention. However, they do not provide precise estimates of effect sizes."

… but …

"All evaluation methods can be considered and used as part of a theory-based approach (particularly research methods that aim to answer process evaluation questions) […]"

The Magenta menu



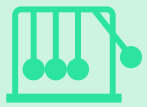
Qualitative Comparative Analysis
Realist evaluation
Process Tracing
Contribution Analysis
Bayesian Updating
Outcome harvesting
Simulation modelling
…

Randomised Control Trial
Propensity Score Matching
Timing of events
Interrupted time series
Instrumental variables
Synthetic Control Methods
Difference in Difference
Regression Discontinuity
…

Flavours of experiment

(Shadish, Cook, and Campbell, 2002, p. 12)

IFF Research



Experiment: “A study in which an intervention is deliberately introduced to observe its effects.” (Almost)

Randomised experiment: “An experiment in which units [e.g., people or schools] are assigned to receive the treatment condition and an alternative condition by a random process…”

Quasi-experiment: “An experiment in which units are not assigned to conditions randomly.”

Examples

IFF Research

Minimum unit pricing on alcohol in Wales	Family Drug and Alcohol Courts	Basic Maths Premium – funding model
Maths Through Picture Books	School-based counselling	Mentoring programme to prevent knife crime

Theory-based evaluation – in the good old days

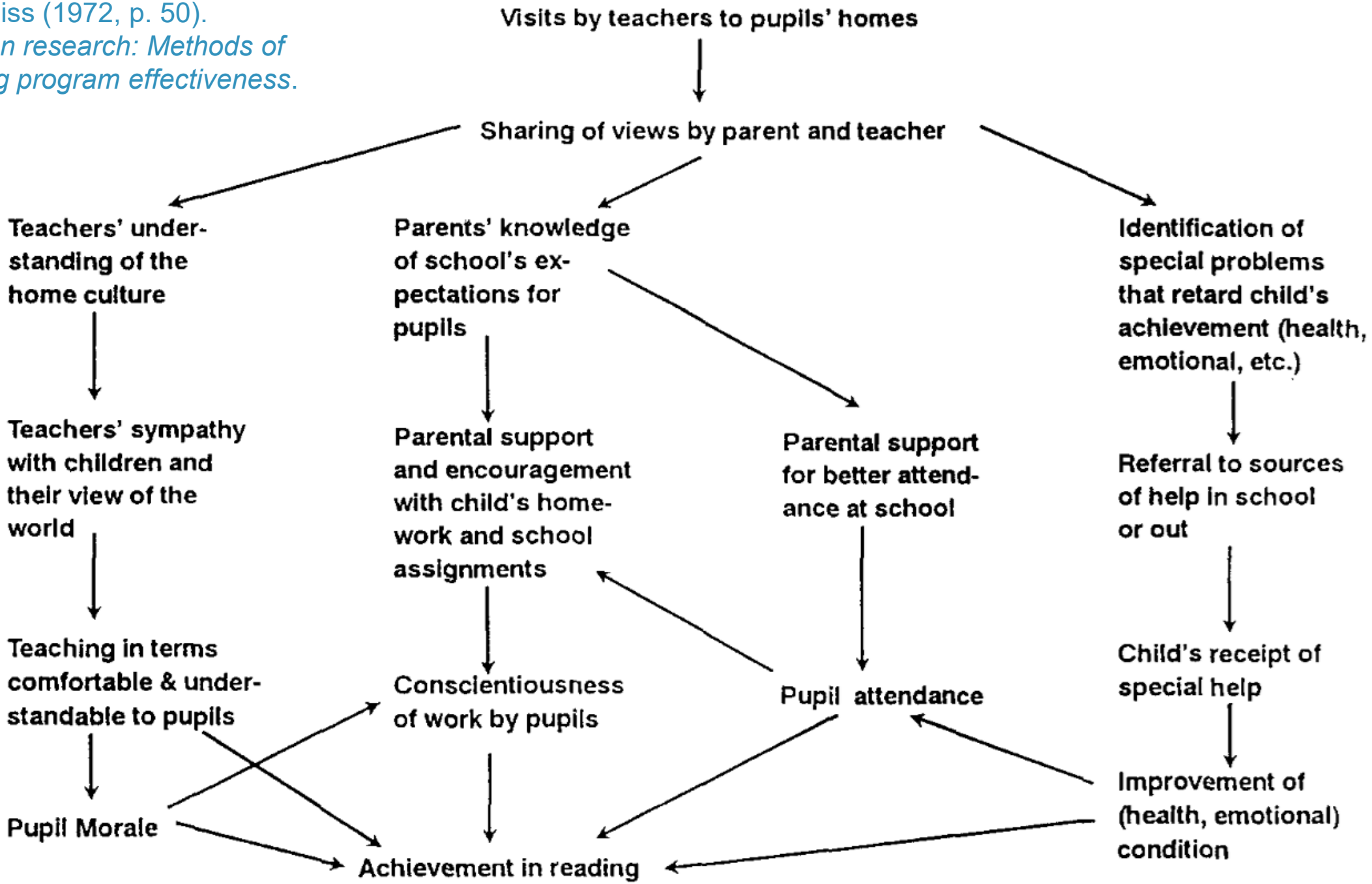
IFF Research

"A theory-based evaluation of a program is one in which the selection of program features to evaluate is determined by an explicit conceptualization of the program in terms of a theory […] which attempts to explain how the program produces the desired effects."

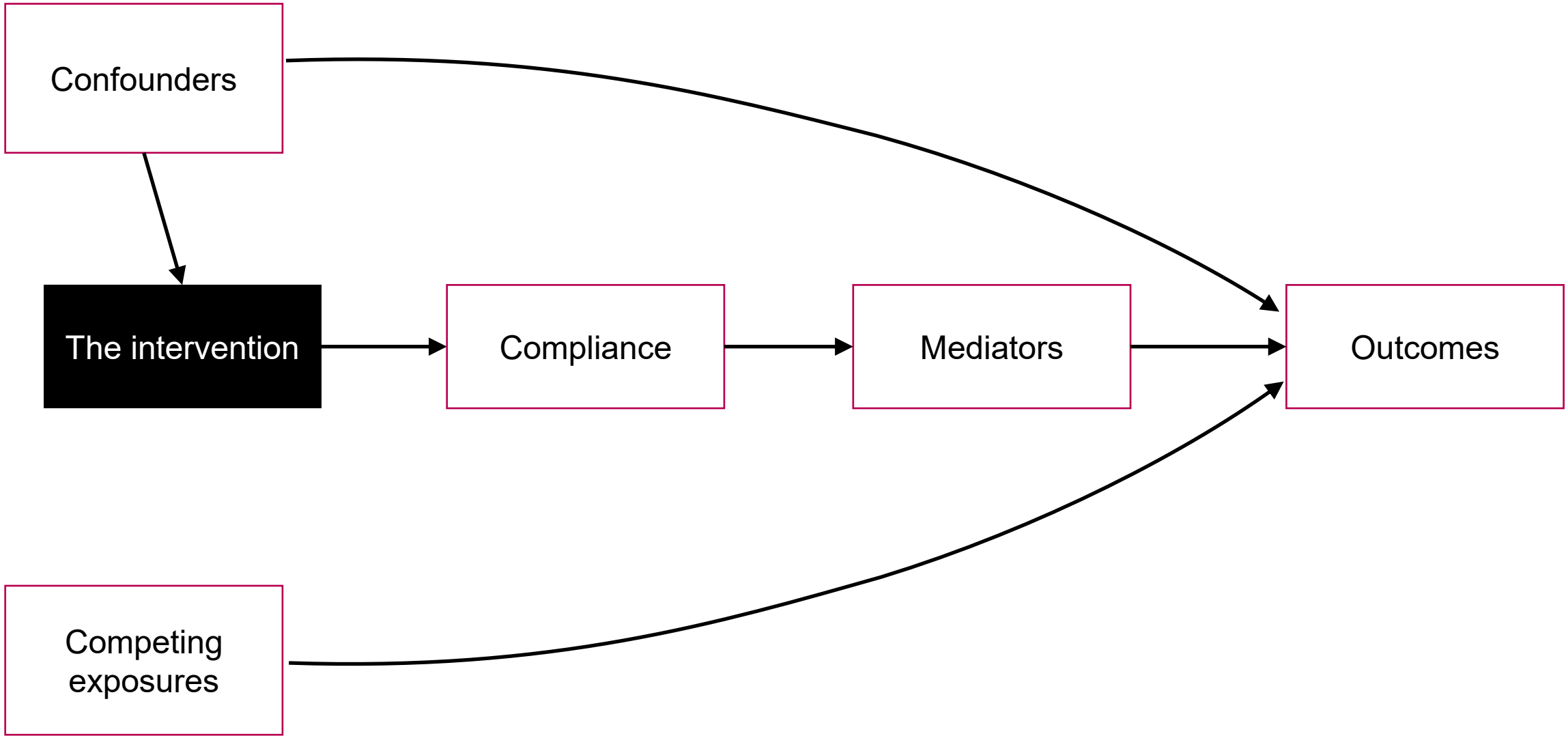
"The essential characteristic is that the theory points out a causal relationship between a process A and an outcome B. A → B; that is, A leads to, or causes B."

Carol Taylor Fitz-Gibbon & Lynn Lyons Morris (1975). Theory-based evaluation. *Evaluation Comment*, *5*(1), 1–4.

Carol Weiss (1972, p. 50).
Evaluation research: Methods of assessing program effectiveness.

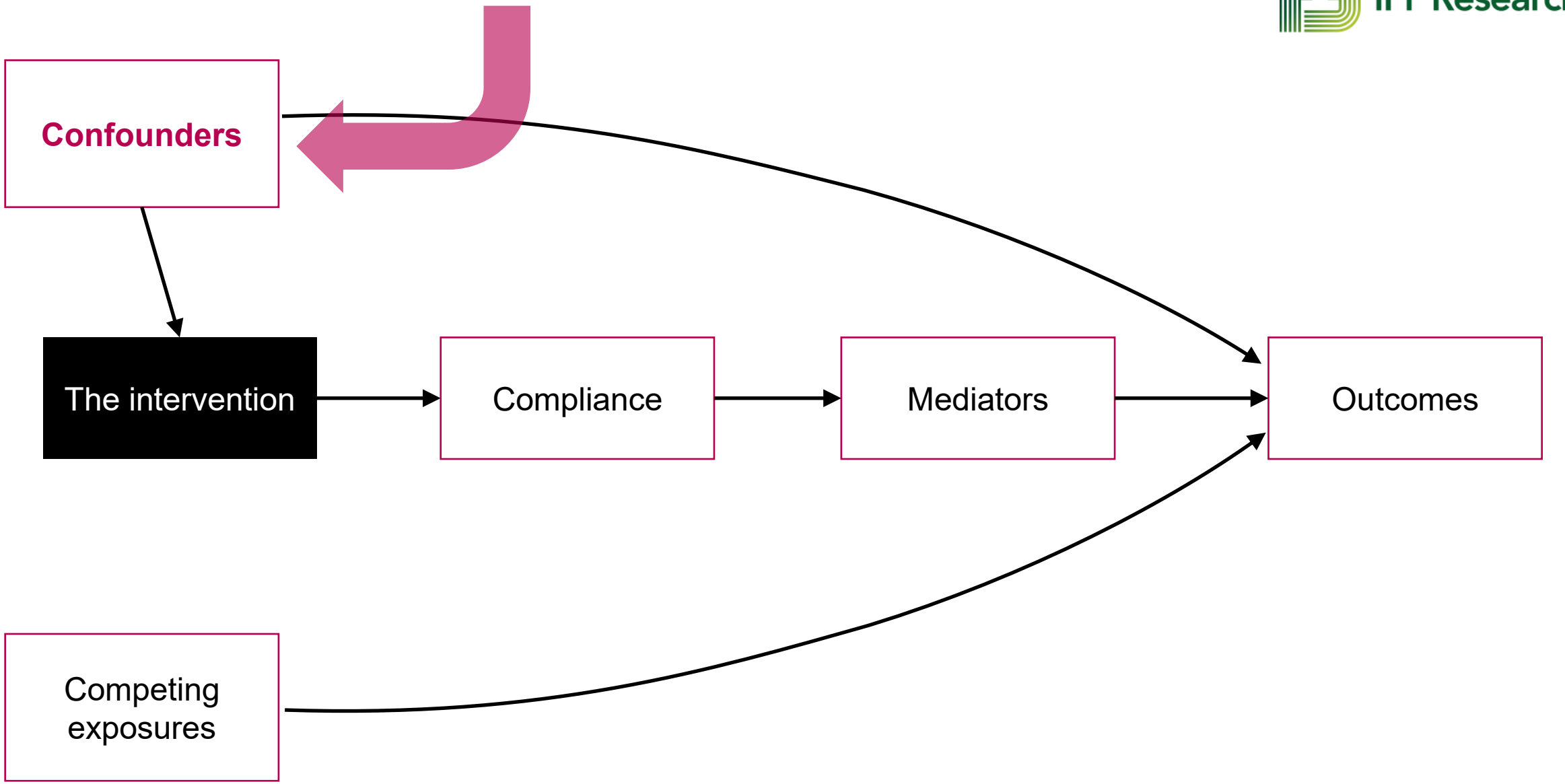


Key features of an experiment



Key headache of a quasi-experiment

IFF Research



LIVES NOT KNIVES

Educate is our school's programme designed to foster a whole-school approach to tackling youth violence and knife crime. It consists of two key initiatives: one focused on mentoring and the other on raising awareness and providing education.

Kerr, J., Morgan, H., MacNaboe, L., Sciarra, A., Cossu, E., Barton-Crosby, J., & Fugard, A. (2023). *LNK Educate: Feasibility and pilot study report*. Youth Endowment Fund.

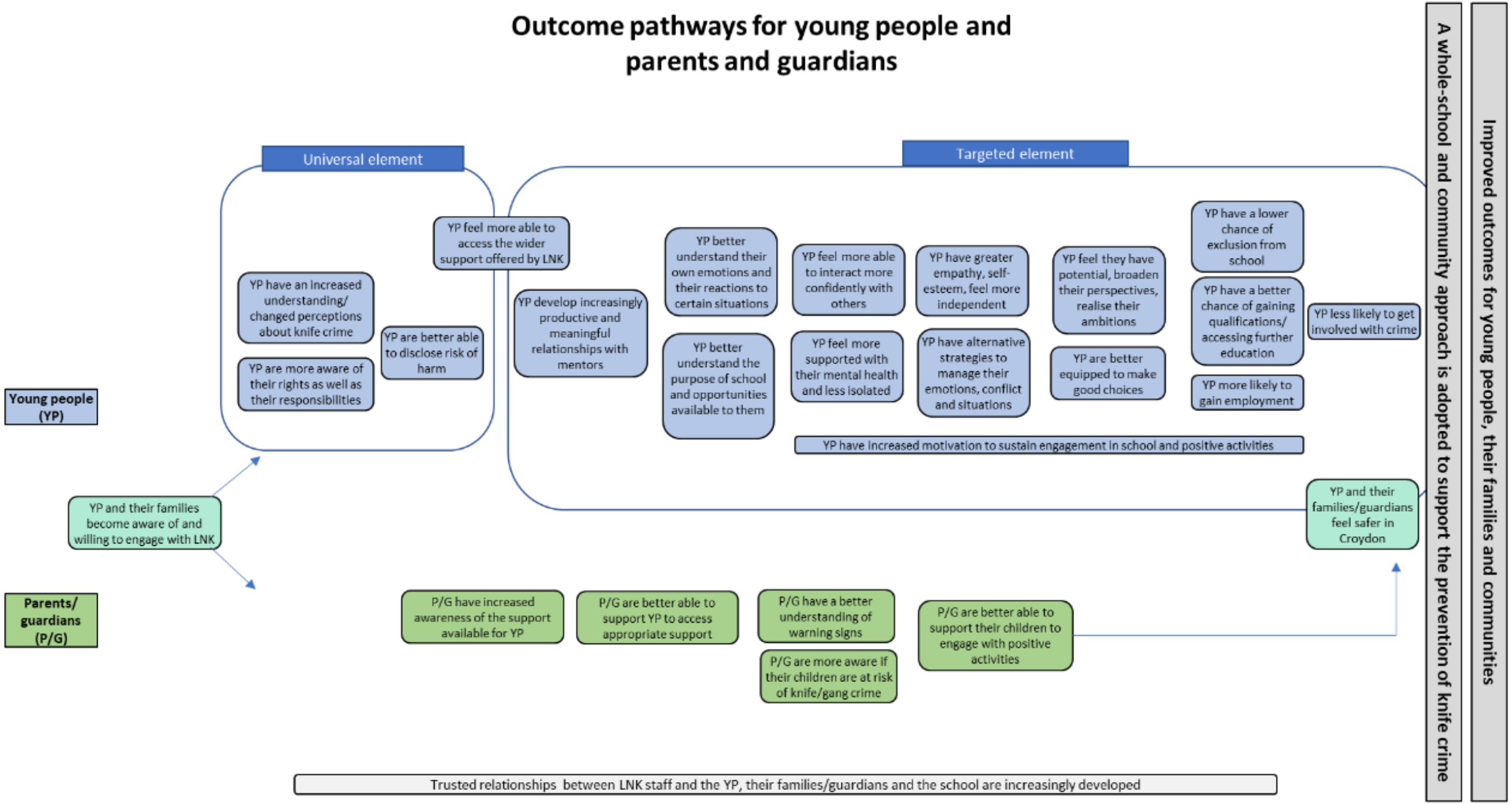
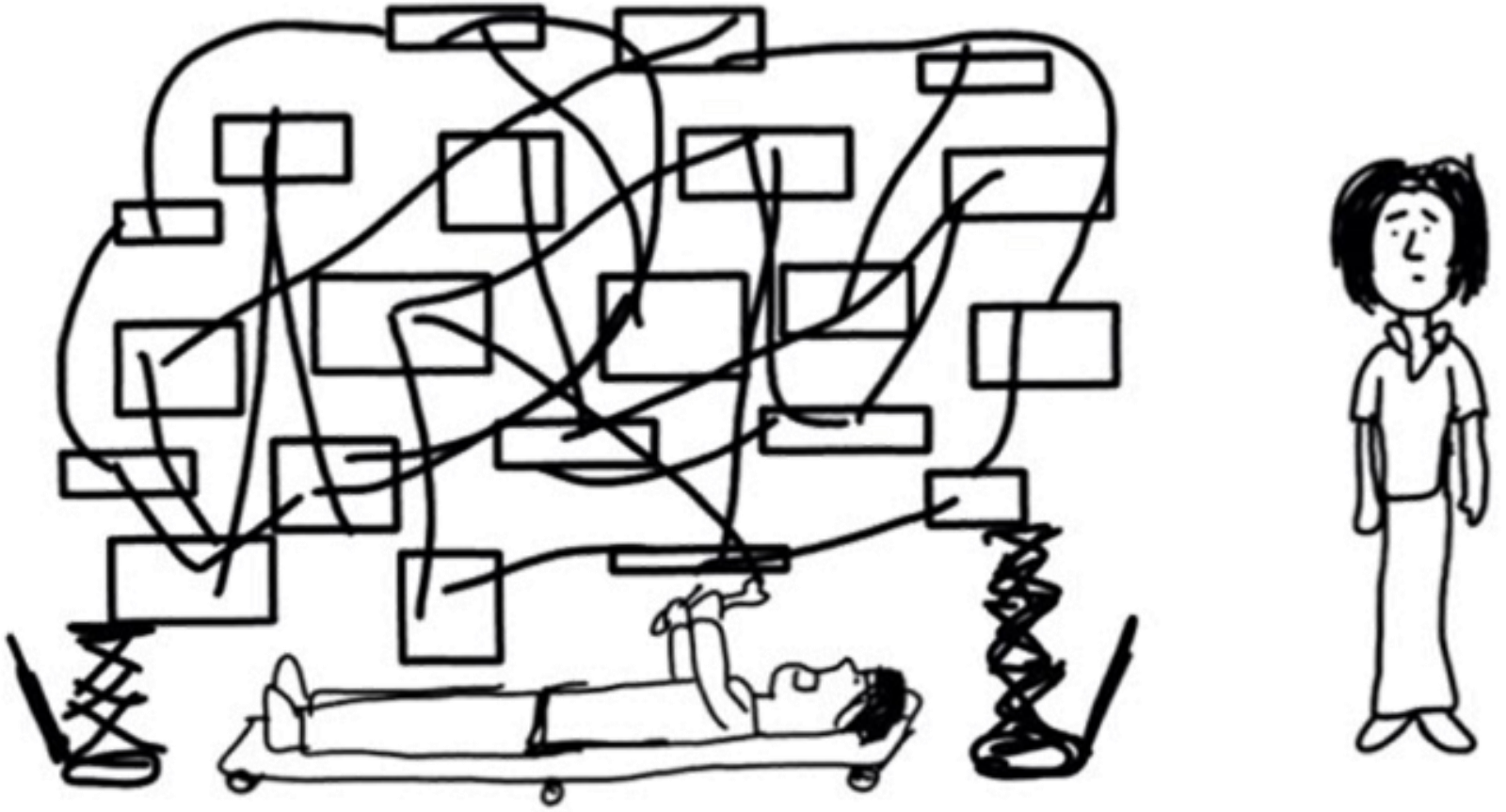


Figure 2. Outcome pathways for young people and parents and guardians

“At the logic model repair shop”



- **Targeted element:** following the intended outcomes of the universal element, the pathway moves on to the intended outcomes for young people who receive the targeted element of LNK Educate. It is expected that attending the mentoring sessions will achieve the following outcomes:

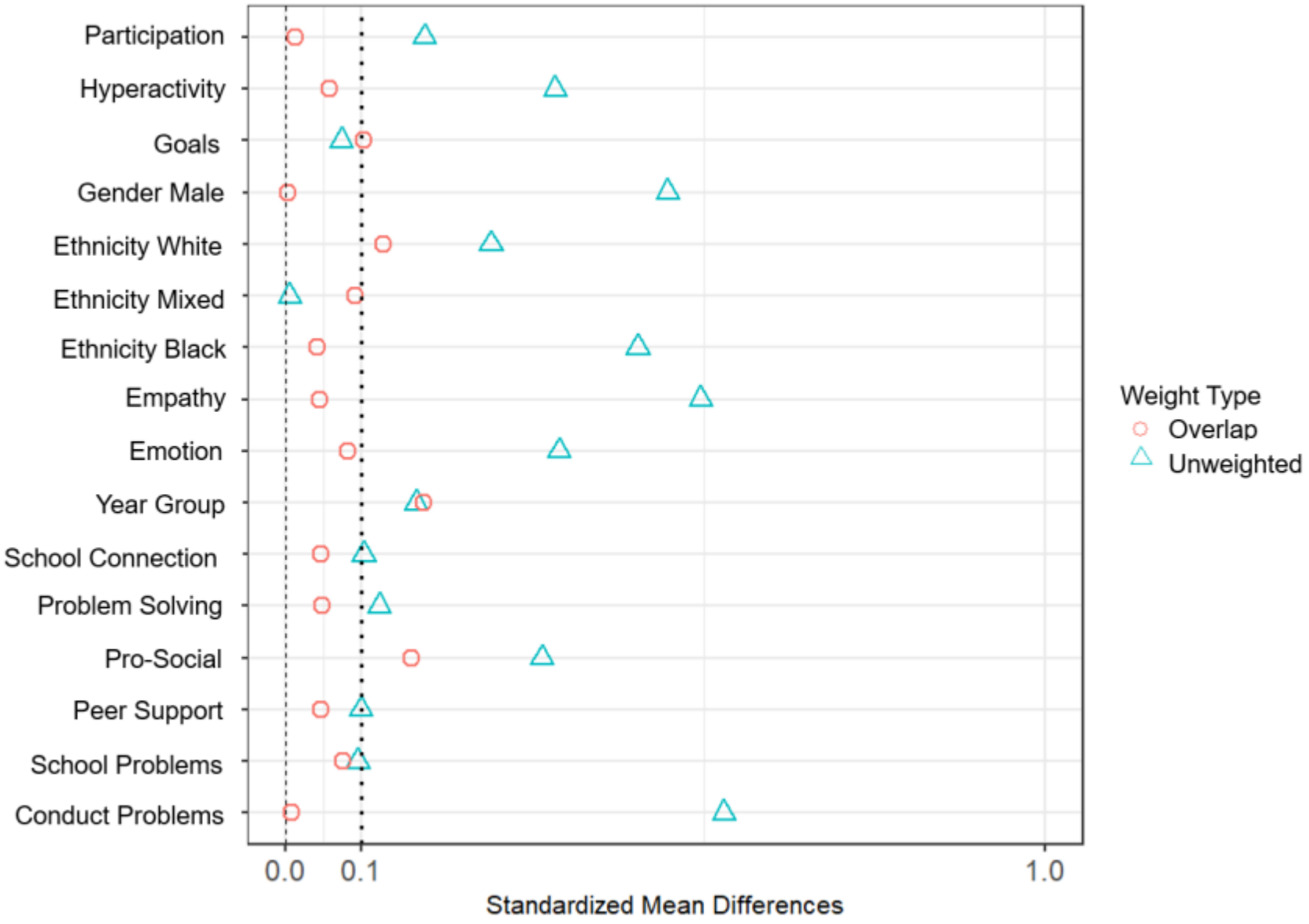
    - As young people embark on, and progress through, the mentoring process, they develop increasingly productive and meaningful relationships with their LNK mentor.

    - It is hoped that the mentoring relationships provide an environment in which young people are able to better understand their emotions (e.g. sadness) and reactions to certain situations (e.g. conflict situations).

    - Simultaneously, the mentoring should result in young people gaining a better understanding of the purpose of school and the opportunities available to them (e.g. good school attainment and career prospects).

    - The combination of improved ability to understand their own emotions and reactions to certain situations and a better understanding of the purpose of school and available opportunities is expected to result in young people feeling more able to interact confidently

You need theory to decide what confounders to measure



Why RCTs could potentially *feel* less theory-dependent

IFF Research

> "The RCT is neat because it allows us to learn causal conclusions without knowing what the possible confounding factors actually are."
>
> – Nancy Cartwright (2010, p. 64)

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*, 59–70.

Why RCTs could potentially *feel* less theory-dependent

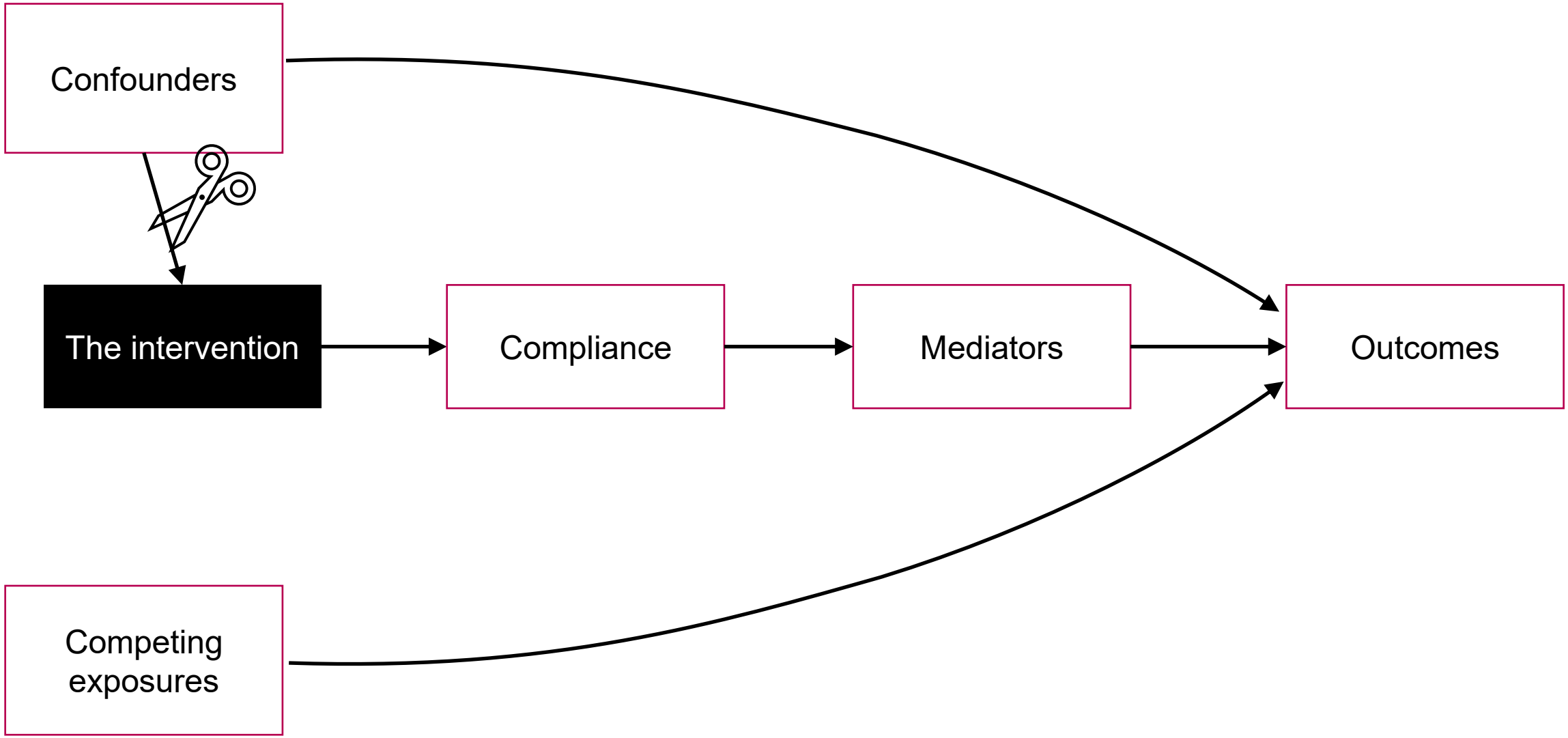
IFF Research

"The RCT is neat because it allows us to learn causal conclusions without knowing what the possible confounding factors actually are."

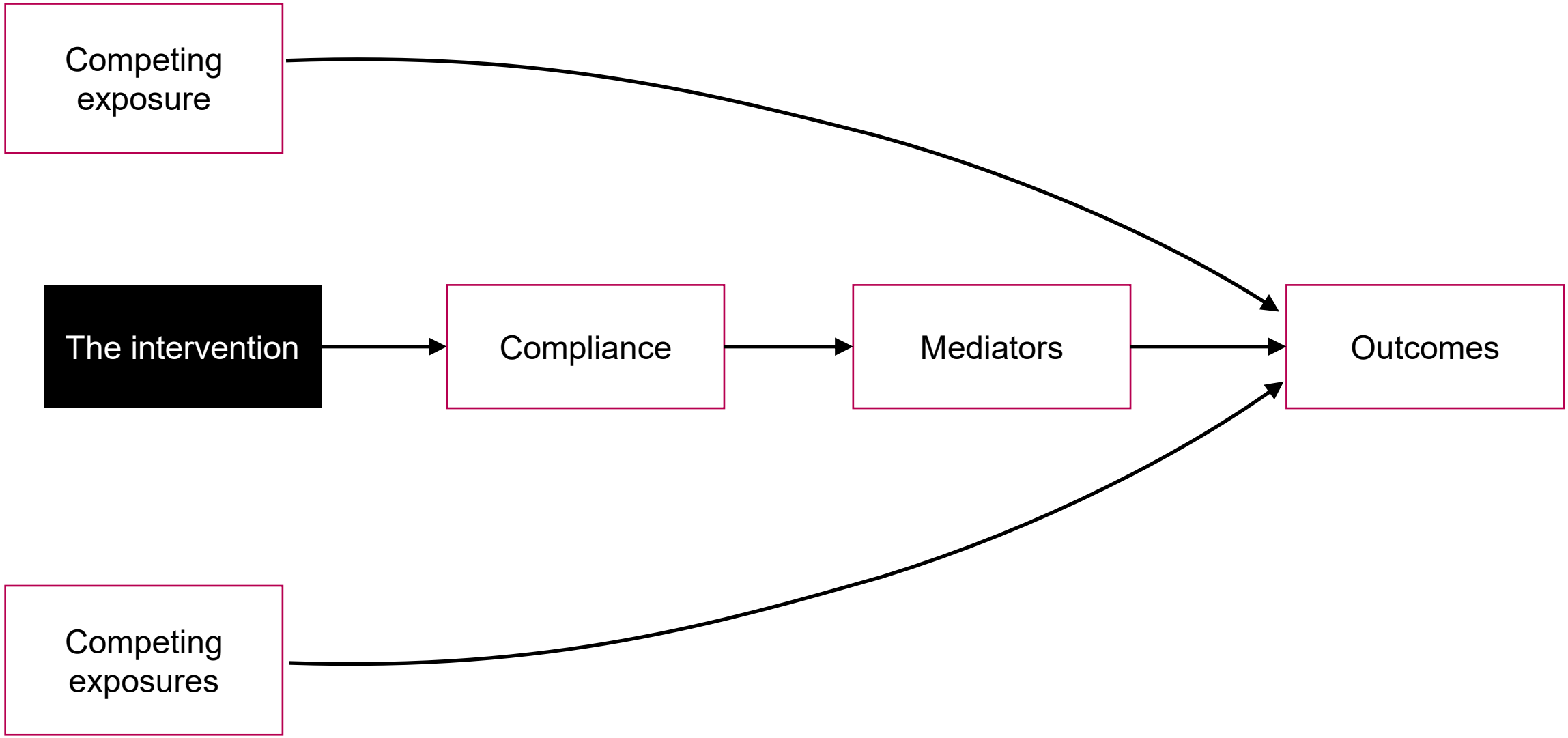
– Nancy Cartwright (2010, p. 64)

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*, 59–70.

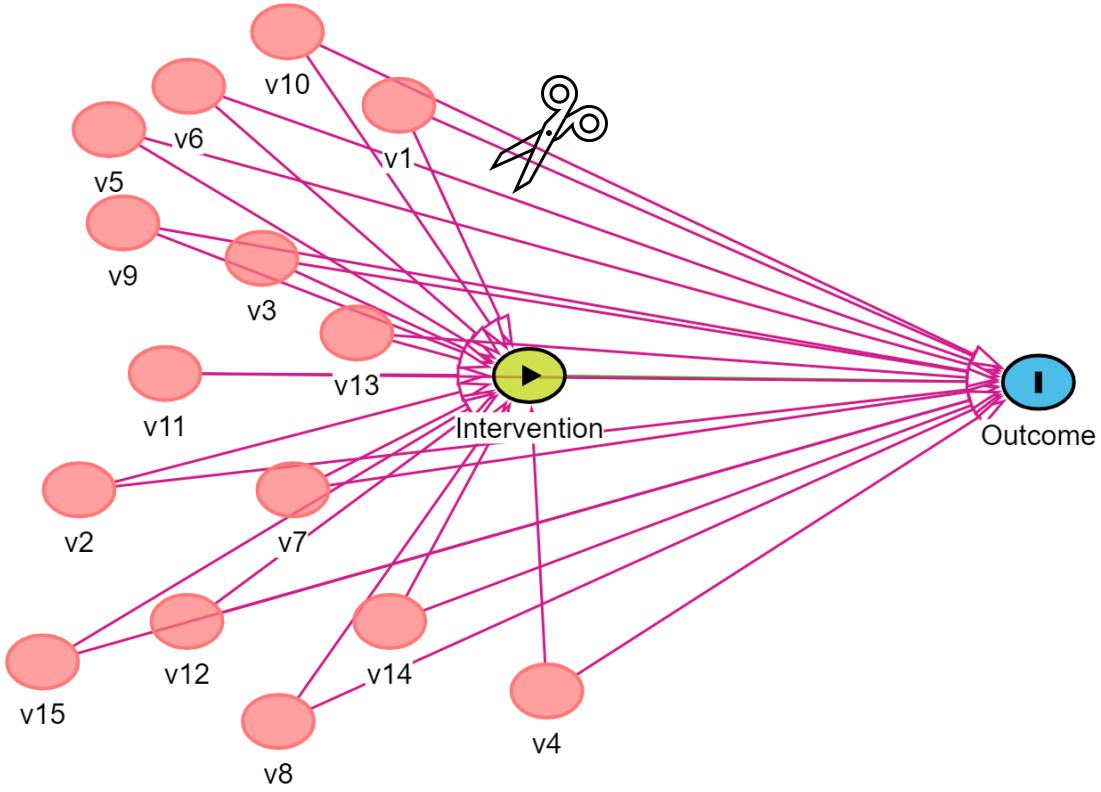
Randomisation turns confounders into competing exposures

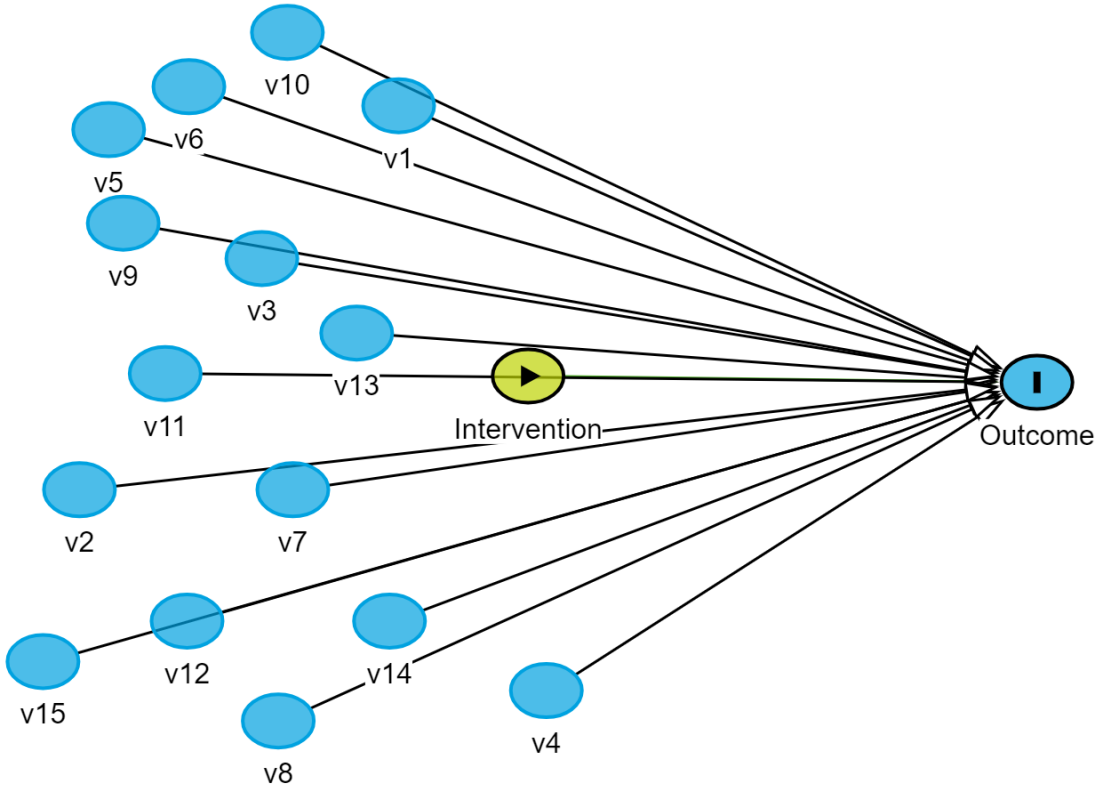


Randomisation turns confounders into competing exposures

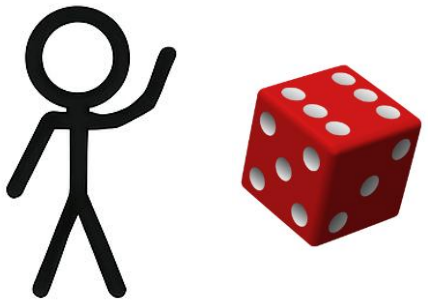


All confounders – even those we can't measure





One obvious place theory enters again



School programme to prevent knife crime

Teaching as usual

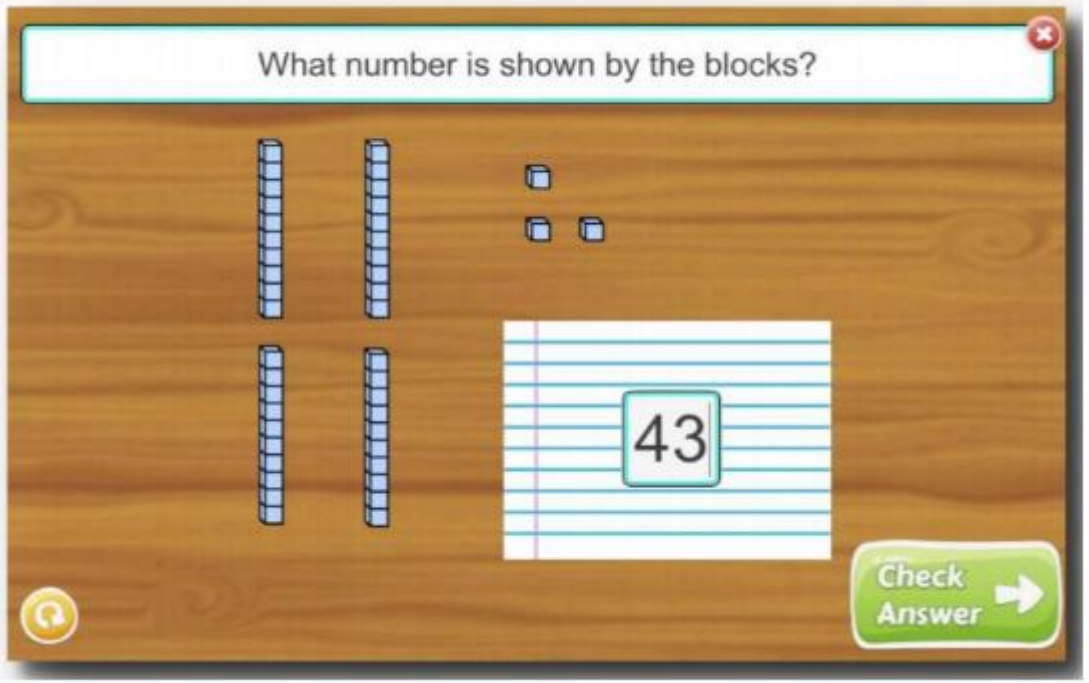
What's the difference in programme content?

Another place – selecting covariates again
(Huey Chen and Peter Rossi, 1983, p. 292)

IFF Research

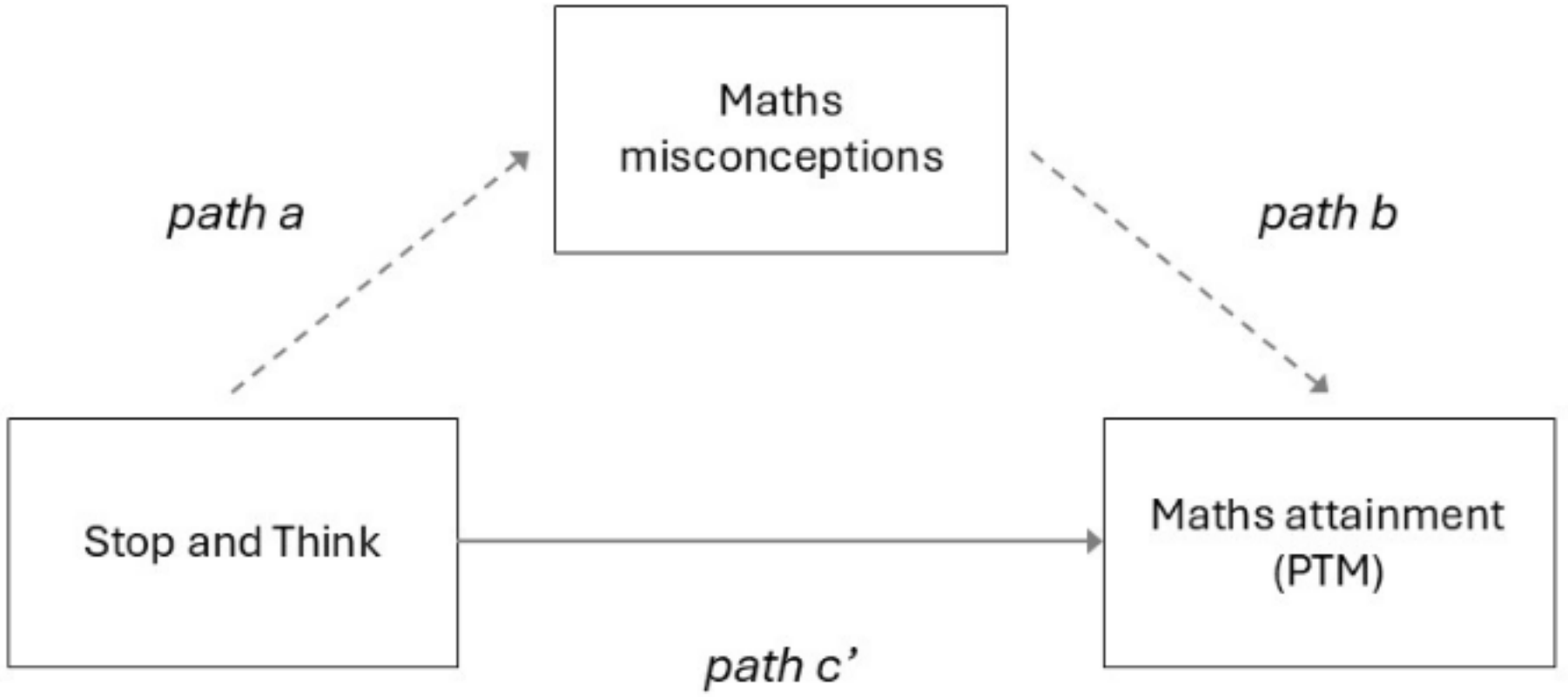
"… advocates of the black box experimental paradigm often neglect the fact that after randomization exogenous variables are still correlated with outcome variables. Knowing how such exogenous factors affect outcomes makes it possible to construct more precise estimates of experimental effects by controlling for such exogenous variables."

Stop and Think



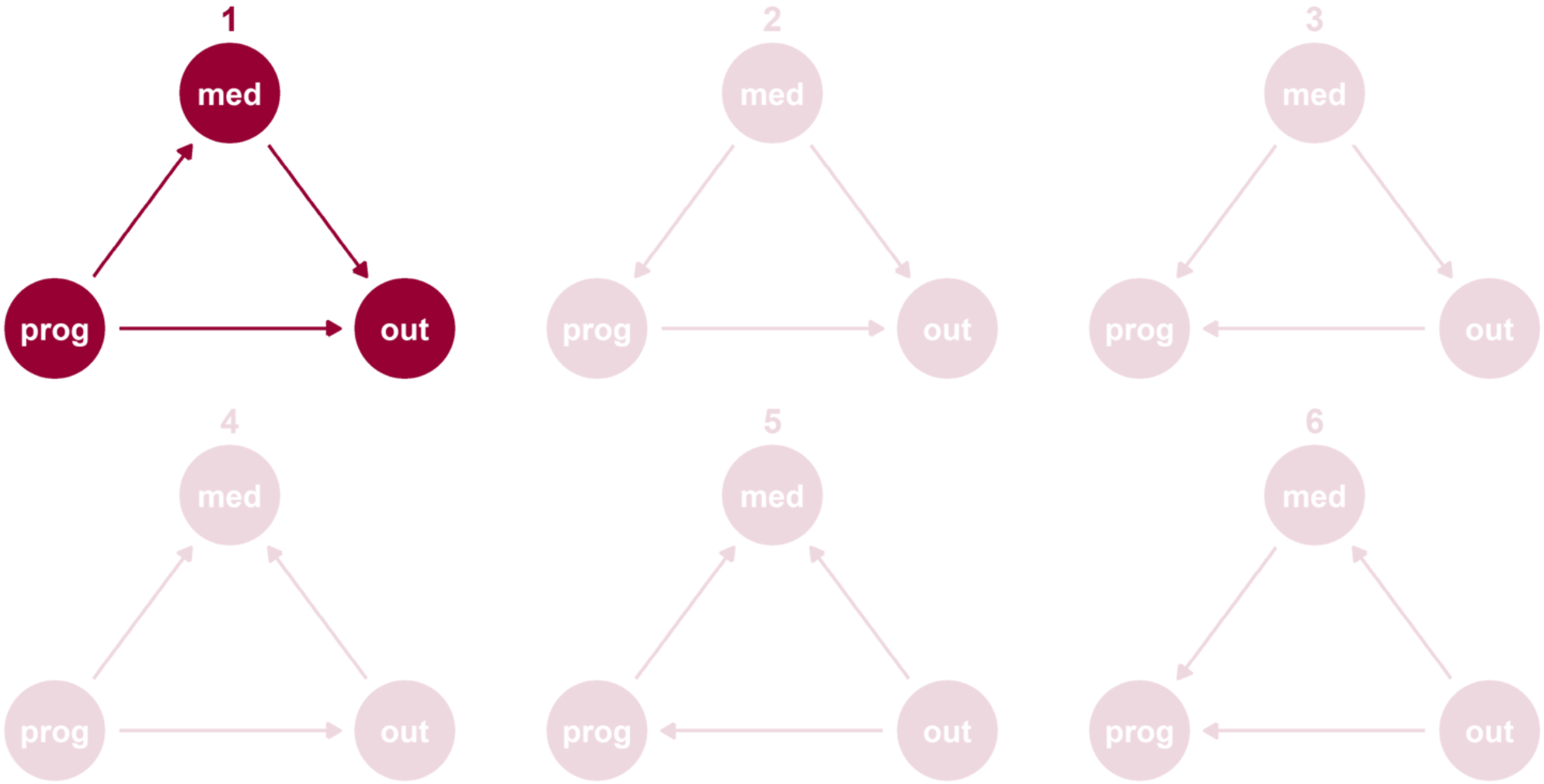
Takala, H., Kuo, T.-L., Duysak, E., Bhatti, S., Stoilova, E., Fletcher, A., McGuinness, N., McKaskill, M., & Fugard A. (2025). Stop and Think: Learning Counterintuitive Concepts Evaluation Report. Education Endowment Foundation.

Another – mediation tests



Even with all the data you need, different causal models are statistically equivalent

IFF Research



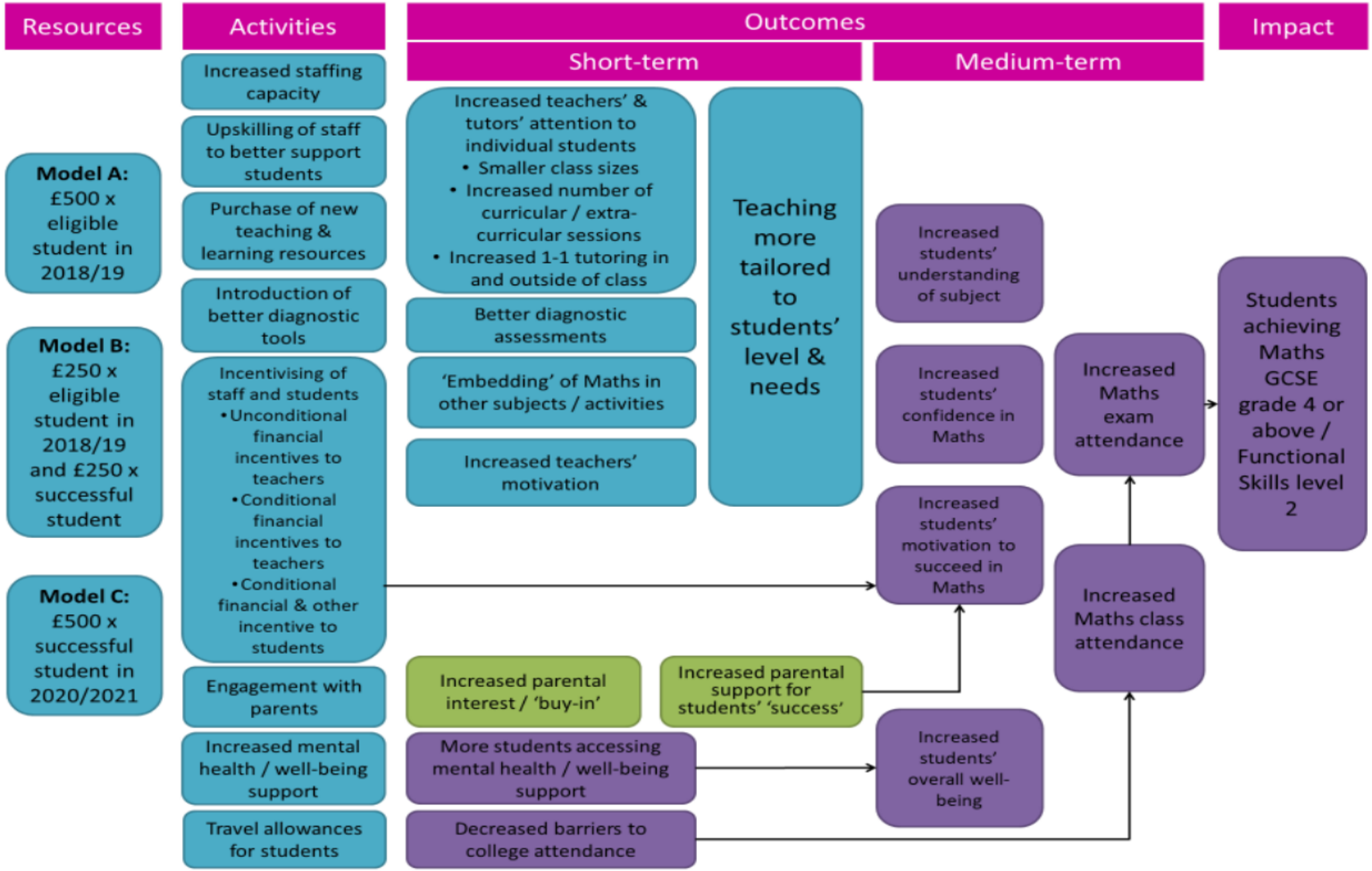
Education Endowment Foundation

Basic Maths Premium Pilot

Evaluation Report

February 2024

Molly Scott, Berenice Scandone, Julia Griggs, Emily Roberts, Tom Bristow, Eleanor Woolfe, Monica Dey, Andi Fugard


Resources
Activities
Outcomes
Short-term
Medium-term
Impact
Model A: £500 x eligible student in 2018/19
Model B: £250 x eligible student in 2018/19 and £250 x successful student
Model C: £500 x successful student in 2020/2021
Increased staffing capacity
Upskilling of staff to better support students
Purchase of new teaching & learning resources
Introduction of better diagnostic tools
Incentivising of staff and students
• Unconditional financial incentives to teachers
• Conditional financial incentives to teachers
• Conditional financial & other incentive to students
Engagement with parents
Increased mental health / well-being support
Travel allowances for students
Increased teachers' & tutors' attention to individual students
• Smaller class sizes
• Increased number of curricular / extra-curricular sessions
• Increased 1-1 tutoring in and outside of class
Better diagnostic assessments
'Embedding' of Maths in other subjects / activities
Increased teachers' motivation
Teaching more tailored to students' level & needs
Increased parental interest / 'buy-in'
Increased parental support for students' 'success'
More students accessing mental health / well-being support
Decreased barriers to college attendance
Increased students' understanding of subject
Increased students' confidence in Maths
Increased students' motivation to succeed in Maths
Increased students' overall well-being
Increased Maths exam attendance
Increased Maths class attendance
Students achieving Maths GCSE grade 4 or above / Functional Skills level 2

Also, trials and quasi-experiments usually include qualitative research too

‘The fact that we’ve got this potential funding in the future that will reward us for that, that’s great, but we’ve still got to find the funds now to do what we do’
(Head of maths, Model C).

‘I think it’s really unfair because you would not really know how much money you were going to get. I don’t think we would have been able to spend any additional money on that basis, so for us it wouldn’t have really worked. ... I would never have been able to employ two staff on the basis that I might get a certain amount of students through a GCSE. That would have been too much of a financial risk’
(Head of maths, Model A)

Also, trials and quasi-experiments usually include qualitative research too

‘The fact that we’ve got this potential funding in the future that will reward us for that, that’s great, but we’ve still got to find the funds now to do what we do’
(Head of maths, Model C).

‘I think it’s really unfair because you would not really know how much money you were going to get. I don’t think we would have been able to spend any additional money on that basis, so for us it wouldn’t have really worked. ... I would never have been able to employ two staff on the basis that I might get a certain amount of students through a GCSE. That would have been too much of a financial risk’
(Head of maths, Model A)

So what…?

All the things are theory-based if you read closely enough.

However, this confusion is persistent:

- How invitations to tender
- How pros and cons of different approaches are discussed
- How proposals are written
- The narrative in reports
- Debates in the field

Together, it actively impedes thinking

Muddle 2:
"counterfactual" is a synonym for "comparison group"

What is a counterfactual?

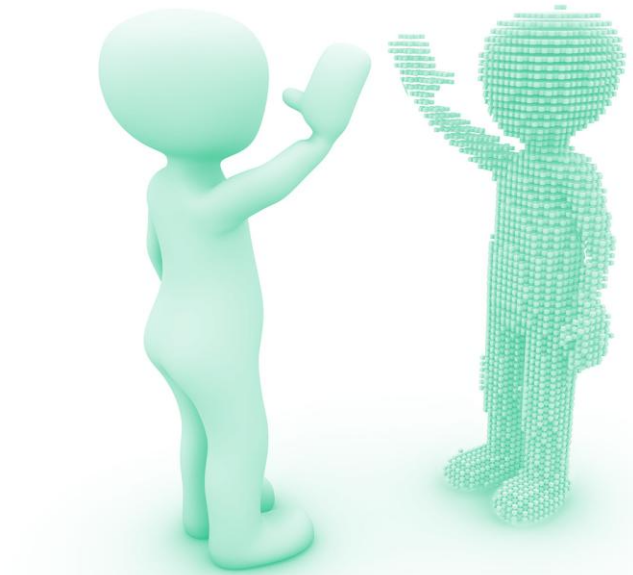
Counterfactual conditional	1. If I had got out of bed 10 mins earlier, I wouldn't have missed the train 2. If Jane had been in the comparison group, her maths test result would have been lower
Counterfactual query	1. If I had got out of bed 10 mins earlier, would I have missed the train? 2. If everyone in the intervention group had instead been in the comparison group, what would their average score have been?

Individual causal effects (e.g., Neyman–Rubin)

IFF Research

Actual outcome

Actual outcome following the introduction of a programme



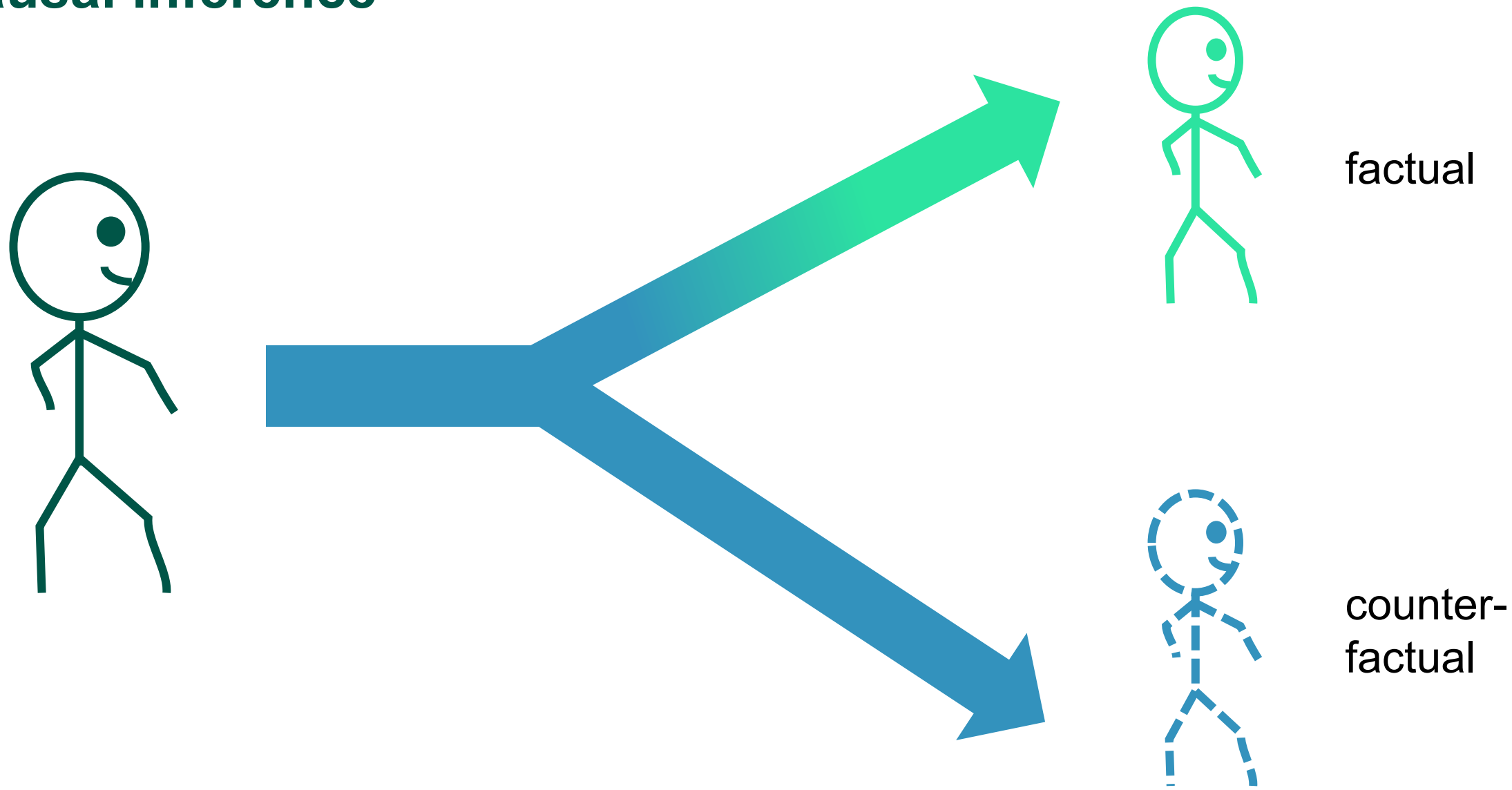
Counterfactual outcome

What the outcome would have been if, e.g., the programme had not been introduced

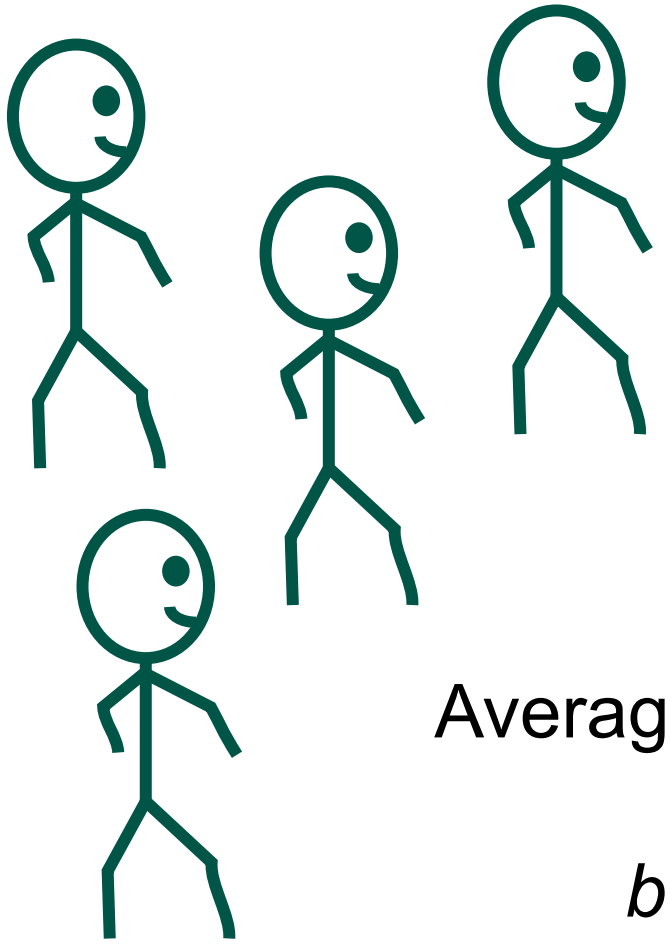
An **individual's causal effect** is the difference

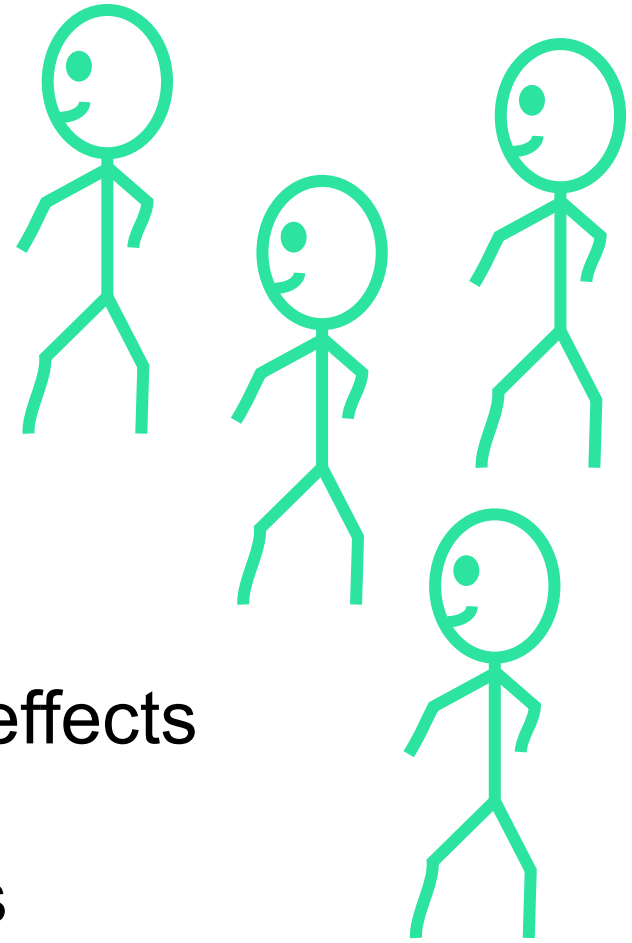
The “fundamental problem” of causal inference

IFF Research



Counterfactual outcomes can be estimated statistically





Average of *within*-person causal effects
estimated using
between-person differences

Example assumptions

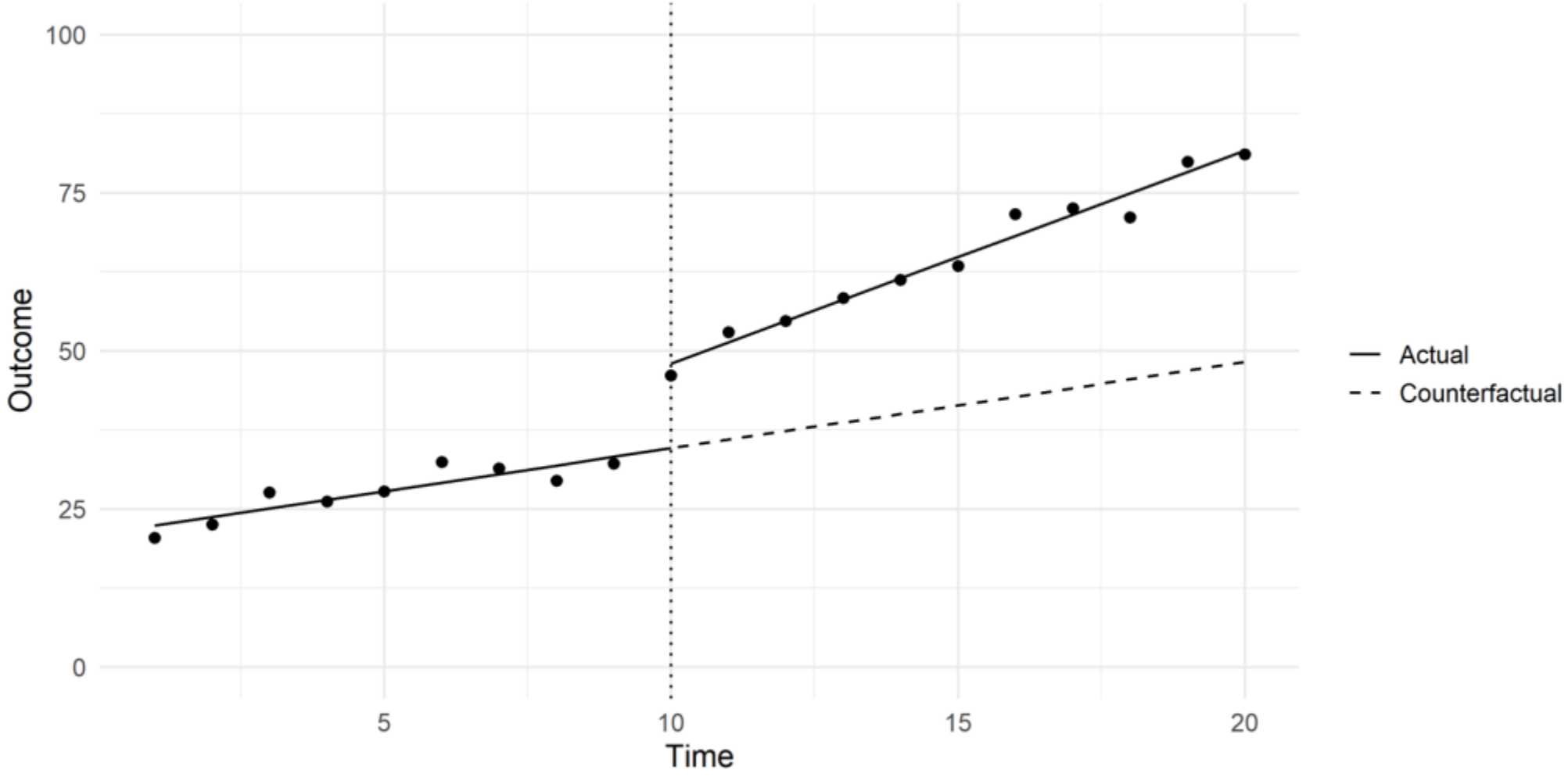
Sufficiently well-defined intervention and control: we need a precise definition of what we're contrasting

Relevant outcomes data: we need measures of outcomes following conditions that mirror the definitions

No unmeasured confounding: relevant pre-intervention measures and they are statistically adjusted
(or there is random assignment)

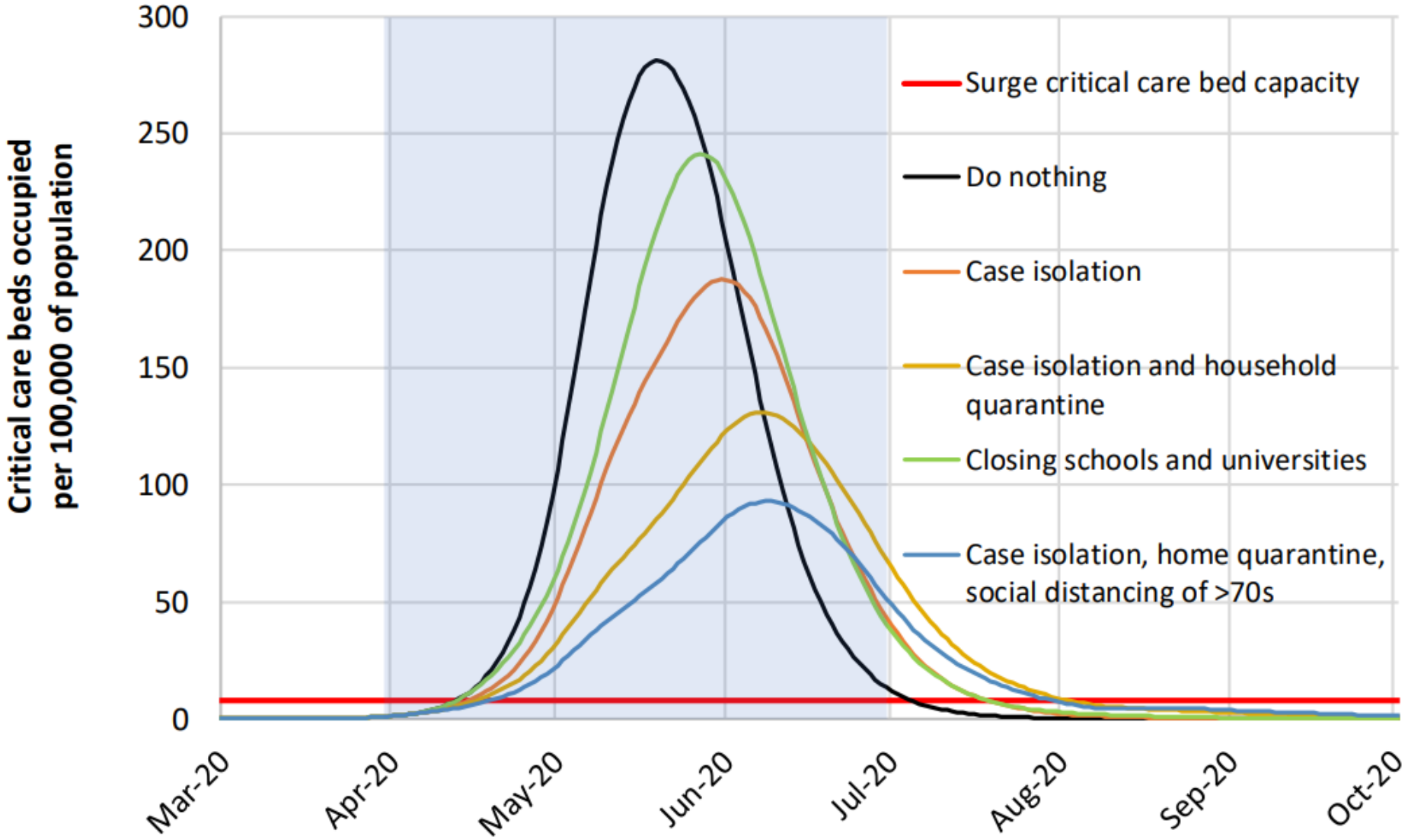
Statistical counterfactuals can be estimated without a comparison group

IFF Research



When we know lots about a topic area, we can simulate counterfactuals

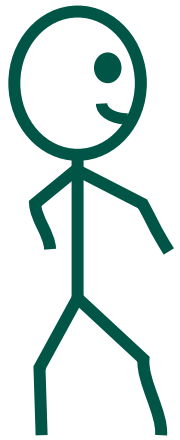
IFF Research



Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand

There are also mental models of counterfactuals

If we hadn't received that grant, the community centre would have had to close

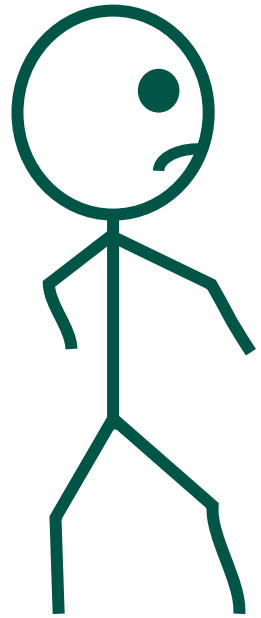


The grant was helpful and enabled us to stay open

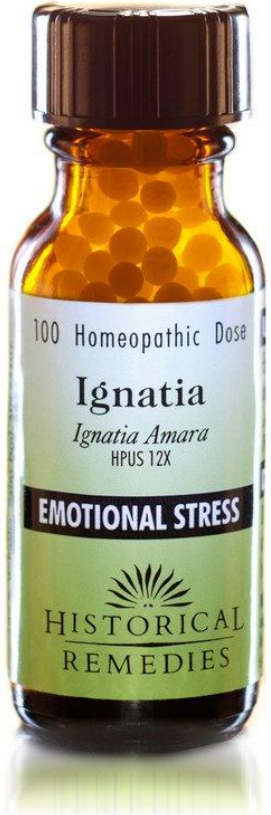
See the extensive psychology of reasoning (e.g., Byrne, 2016; Gerstenberg, 2024; Over & Cruz, 2023)

Sometimes it's hard

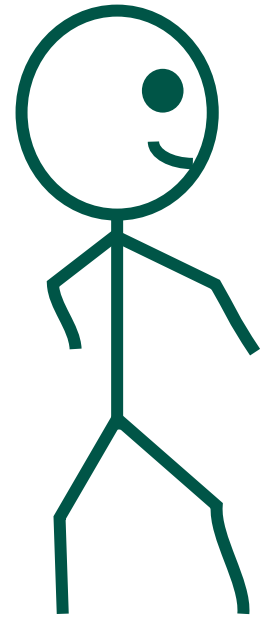
IFF Research



Feeling (unusually) low after an (unusually) stressful day



Friend suggests a homeopathic remedy

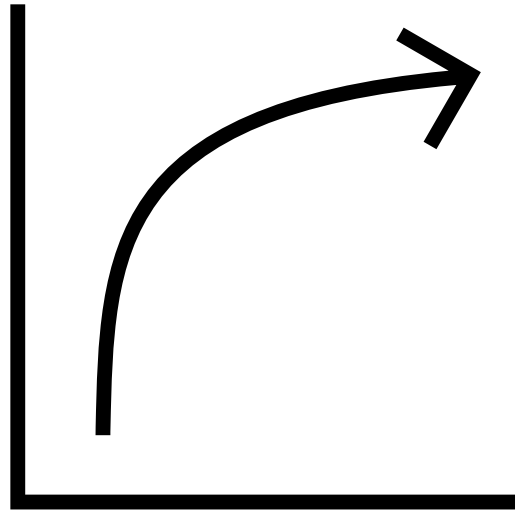


Feeling back to normal after a couple of days

Sometimes less challenging

Family living in bed poverty for over a year

Charity transforms child's bedroom



Child is healthier and happier

Based on example in Bed poverty charity's ultimate aim 'to not exist'

Counterfactuals appear in policy appraisal – just with different terminology (Green Book)

IFF Research

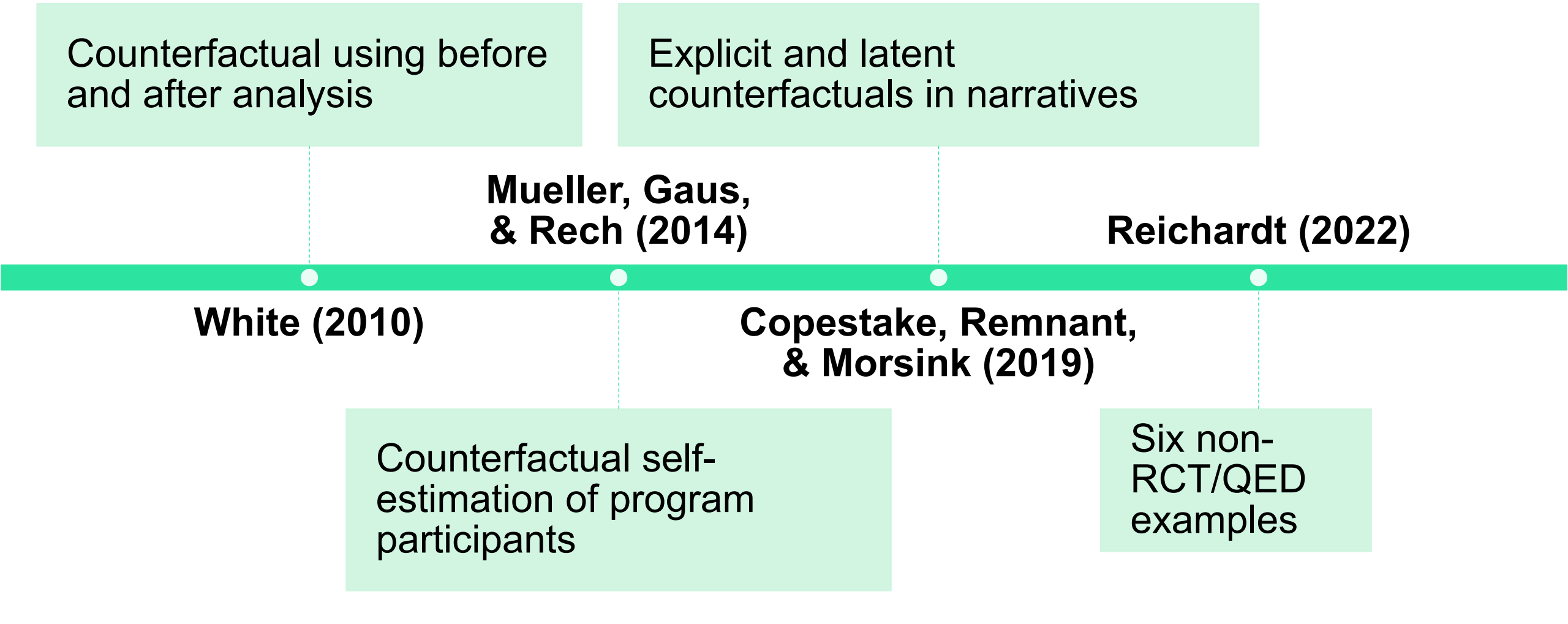
4.17 Fully understanding the BAU helps practitioners to distinguish between deadweight and additionality. These terms are defined as follows:

- **Deadweight:** Outcomes that would have taken place without any intervention. For example, a proposal to achieve growth in a particular industry should consider the market trends that would occur regardless of government intervention.
- **Additionality:** Outcomes that take place as a result of the intervention. For example, a proposal to achieve growth in a particular industry should consider how it is genuinely raising productivity over and above market trends.

4.18 Practitioners should ensure that their description of BAU is fair and reasonable. Overestimating the costs and risks of BAU will make government action appear unduly attractive. Underestimating them will make government action appear unduly unattractive.

Evaluation has also been pondering the full breadth of counterfactual evidence for some time

IFF Research



Case study: contribution analysis

IFF Research

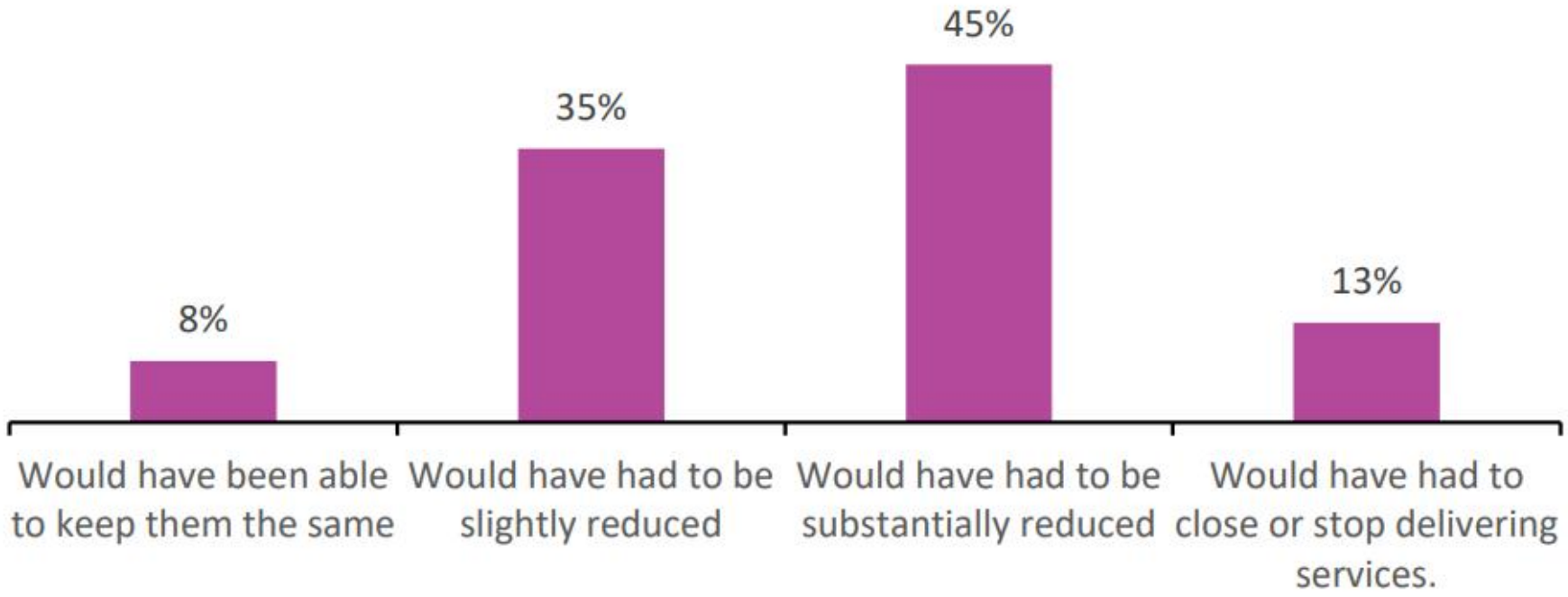
> “In many situations a counterfactual perspective on causality […] is unlikely to be useful; experimental designs are often neither feasible nor practical.”
>
> “[Contribution analysis] uses a stepwise (generative) not a counterfactual approach to causality.”

Mayne (2019, pp. 173–4, 176)

Part of a contribution analysis
(Rosa Lau et al., 2022, p. 44)

IFF Research

Figure 4:6 Ability to maintain the same level of service delivery without the VCSE sector funding



A bunch of counterfactuals

Critical appraisal – example considerations

Statistical counterfactual	Self-reported counterfactual
Check testable assumptions, such as baseline balance	Check whether narrative explanation is coherent
Draw on subject matter expertise to determine whether all relevant data was included in the formal model	Draw on subject matter expertise to determine whether all relevant data was included in informal argument

So what…?

Instead of “there is no counterfactual, so we use X”

→What is the counterfactual?
→How might be develop an argument for it
→What fields like the psychology of reasoning could we draw on
→What are the risks of bias in a particular policy area

This makes it easier to think about designs

Muddle 3:
"realist" evaluation is a distinct kind of evaluation

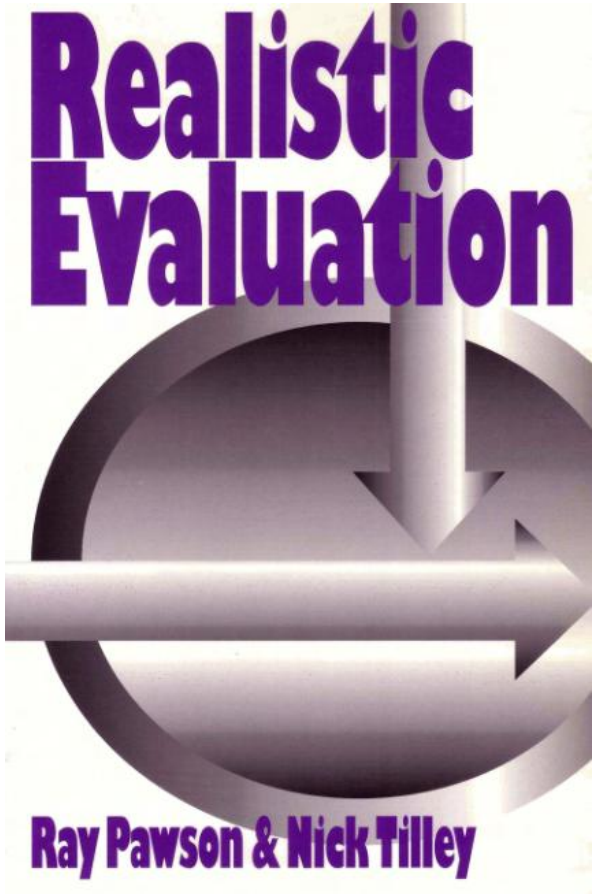
What is realist evaluation?

IFF Research

"Realist Evaluation seeks to understand how a policy or programme causes the desired outcomes. The purpose of a realist evaluation is as much to test and refine the theory behind the policy intervention as it is to determine the policy's outcomes in a particular set of circumstances. A realist evaluation asks not 'what works?' or 'by how much does it work' but 'what works, for whom, in what respects, to what extent, in what contexts, and how?'"

Magenta Book 2020 – Realist Evaluation

Where it came from (1997)

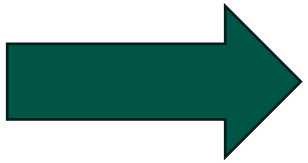


- Suggests how to apply scientific methods to policy
- Introduces a bit of philosophy of science
- Alas, something went wrong
- It became an evaluation brand 😬
- Better title:

  Applying scientific thinking to policy evaluation: an introduction

Operant conditioning?

What works best for whom in what context?



Realist evaluation!

Surely that can't be from 1997?

IFF Research

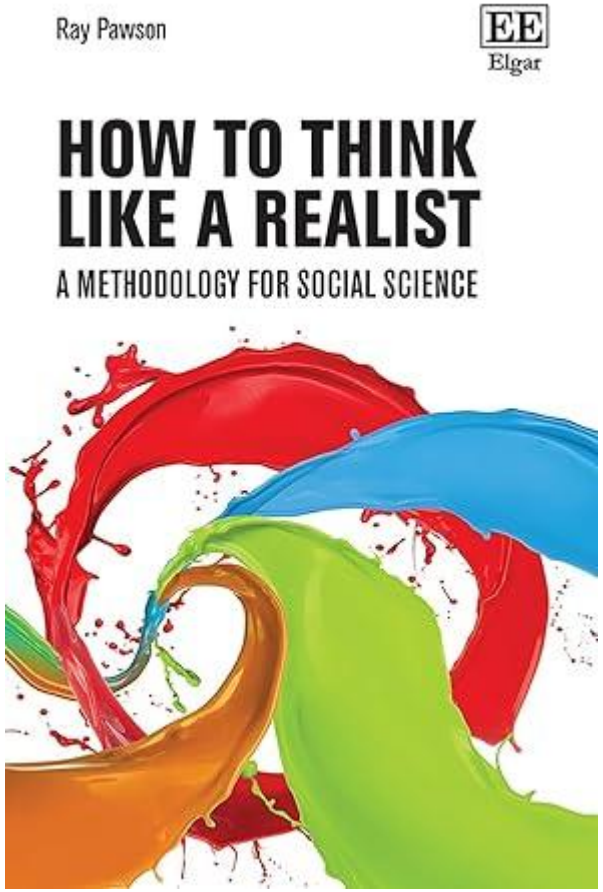
Paul (1967, p. 111):

> […] the question towards which all outcome research should ultimately be directed is the following: *What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?

Pawson and Tilley (1997, p. 10) cite Palmer (1975, p. 150):

> Rather than ask, "What works for offenders as a whole?" we must increasingly ask "Which methods work best for which types of offenders, and under what conditions or in what types of setting?"

An attempt at correction (Pawson, 2024)?



"All *[!!!]* scientific investigation utilises explanations relating mechanisms and contexts to empirical patterns." (p. 42)

"With Harré, I have characterised generative causation in physical science as the analysis of mechanisms, contexts, and regularities (MCR). In clinical research the focus, quite properly, is on mechanisms, contexts, and outcomes (MCO). In social research, it might be wise to begin with the shorthand mechanisms, contexts, and change (MCC)." (p. 48)

"The book is no more and no less than an attempt to codify and formalise existing practices." (p. 251)

Also, something I quite like that has broad application – tables!

Context	New mechanism	Outcome
High numbers of prepayment meters, with a high proportion of burglaries involving cash from meters	Removal of cash meters reduces incentive to burgle by decreasing actual or perceived rewards	Reduction in percentage of burglaries involving meter breakage; reduced risk of burglary at dwellings where meters are removed; reduced burglary rate overall

So what…?

The field is littered with debates about whether an RCT or quasi-experiment can be realist

Again leads to muddled

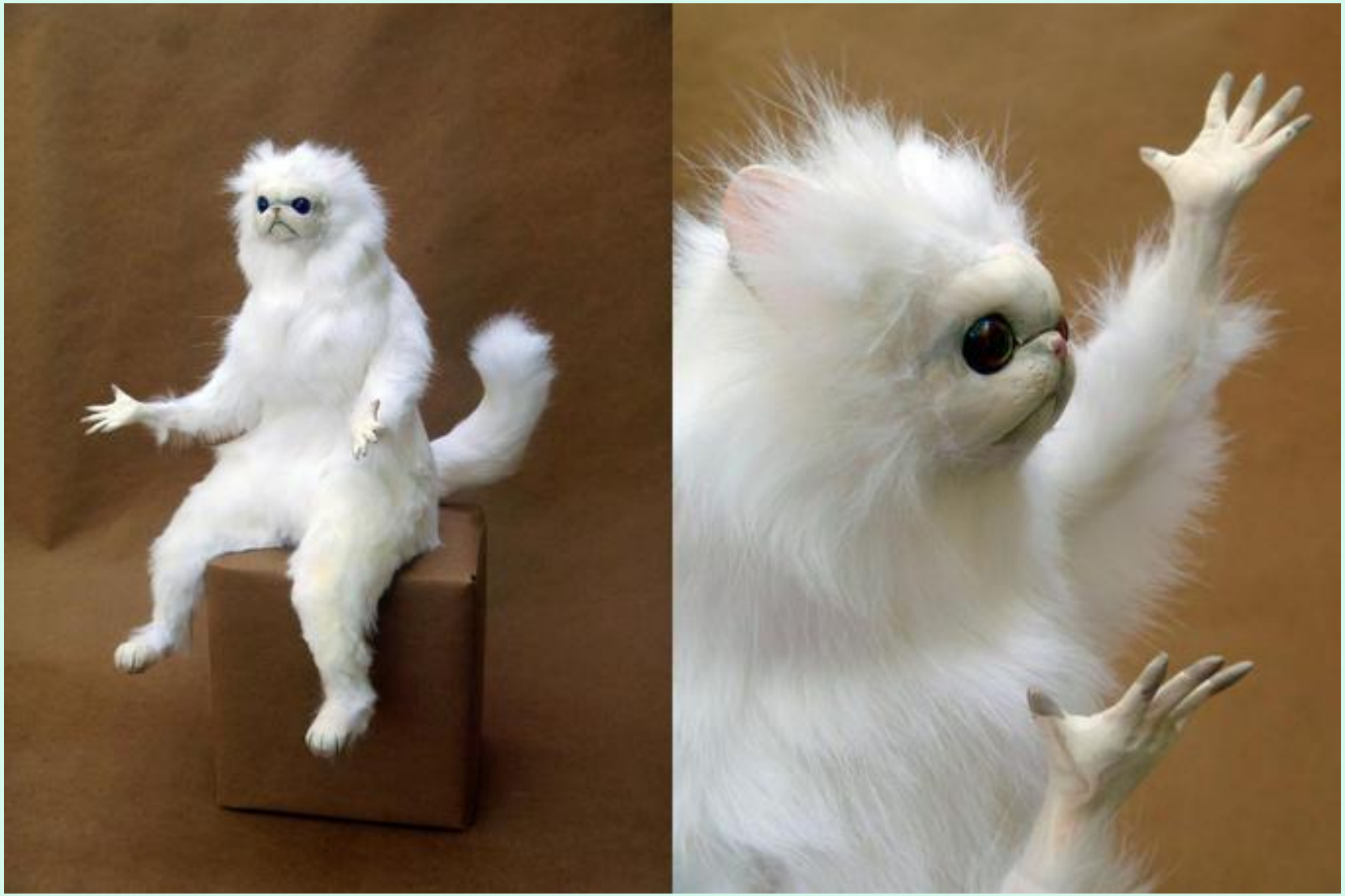
- Invitations to tender
- Reports
- Amateur philosophy everywhere: “reality is real!”
- Historical amnesia – not a good look for a policy profession

Conclusions

It’s apparently easy to get swept up in these muddles – perhaps it’s rewarding?

Challenging artificial divisions between approaches makes it easier to design better studies – easier to think!

Theories are often vague – we can surely do better



Thank you 🙂